UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA VEGETAL

New approaches for taxonomic identification and profiling of polyclonal samples based on Next Generation Sequencing

Rafael Fresca Mamede

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Versão Pública

Dissertação orientada por:
Prof. Doutor Ricardo Pedro Moreira Dias

2018

Abstract

High-throughput sequencing technologies have greatly contributed to developments in the areas of metataxonomics and metagenomics, that encompass the study of complex microbial samples. One of the fundamental steps in the analysis of data derived from the sequencing of such samples is the taxonomic identification of sequencing reads. Despite great availability and diversity of classification tools, taxonomic classification methods deliver variable results with a degree of reliability that is difficult to determine, especially at lower taxonomic ranks, such as species. In this study, three k-mer based classification strategies commonly used to classify metagenomics sequencing data, implemented in Kraken, CLARK and Centrifuge, were applied to metataxonomics simulated data, generated from 16S rRNA gene sequences of reference databases, in order to assess classification performance against an alignment-based strategy typically applied in metataxonomics pipelines. Each classification strategy was evaluated with three distinct reference databases and their condensed variations, and by classifying sequences of 14 subregions of the 16S rRNA gene. Performance was evaluated over 60 and 45 combinations of classifier, database and subregion for the k-mer based strategies and for the alignment-based strategy, respectively. From the k-mer based strategies, the one based on discriminative k-mers stood out and displayed superior performance at the evaluated ranks of family, genus and species. At family-level, classification based on discriminative k-mers was 8.9% (accuracy of 0.878) more accurate than the alignment-based strategy (0.789) and at genus level both strategies achieved the same mean accuracy value (0.704). At species level, classification based on discriminative k-mers displayed superior performance, achieving a maximum accuracy value of ~0.908 and the top 25% of accuracy values obtained with all combinations of classifier, database and subregion were above 0.846. By contrast, the second best strategy, also k-mer based, reached an accuracy maximum of 0.705 and the top 25% of values started at an accuracy value of 0.332. Regarding absolute accuracy values, greater taxonomic resolution was obtained with the classification of sequences from the V4/V3-V4, V4/V3-V4 and V2/V1-V2 subregions for the family, genus and species levels, respectively.

To explore new approaches that might contribute to improvements in classification strategies, the best performing strategy, based on discriminative k-mers, was used to classify sequences from the V2-V4 and V6-V8 subregions. Taxonomic relations between sequences

of both loci were modeled through graph structures. Two graph matching algorithms (maximum weight algorithm and hungarian algorithm) and one simple majority rule algorithm were applied to find matches in each bipartite graph that sought to maximize the similarity between each pair of matched nodes. The application of two cutoff values, determined through the distributions of confidence and gamma values from CLARK's classifications, to a subset of the matches that captured most of discrepancies between taxonomic predictions of both loci caused a drop in the mean accuracy for the V2-V4 subregion, at species level, between 0.73% and 1.99%, depending on the algorithm used to determine matches. For the V6-V8 subregion, an increase in the mean accuracy at species level of 2.48% was achieved for matches determined through the majority rule algorithm. Seven new species, that were not detected in the initial predictions, were detected by altering classification based on the cutoff values, which decreased the number of undetected species from 31 to 24 (in a total of 111 species) when applied to the matches of a simple majority rule algorithm. An alternative approach classified the sequences represented by nodes in the same matches through BLAST. The classification of this subset of sequences improved the mean accuracy at species level for the V2-V4 subregion between 1.12% and 1.53%, depending on the algorithm used to determine matches. Accuracy values for the V6-V8 subregion increased between 5.58% and 7.13%. Using the bitscore of each BLAST prediction to favor high scoring predictions in each match and seek further improvements, led to a slight decrease in the accuracy values for both subregions. Reclassifying a subset of the sequences with BLAST reduced the number of undetected species from 31 to a minimum of 13 and 16, for the hungarian and the majority rule algorithms, respectively. Altering classification based on BLAST's bitscore allowed the detection of two and six additional species, for the hungarian and majority rule algorithms, respectively. Combining sequences from both loci according to the matches determined with the hungarian algorithm or the majority rule algorithm led to multilocus groups in which the constituent sequences shared all taxonomic terms with a mean accuracy of 0.925 and 0.937, respectively.

The results presented in this dissertation demonstrate that a k-mer based strategy developed for metagenomics data may achieve superior performance in the classification of metataxonomics data and that the potential of that strategy can be further explored. In conclusion, these results show that the combination of classification results from two loci and two distinct classifiers constitutes an approach that can bring a considerable improvement to taxonomic classification. This work hence lays the foundation for the development of a strategy that might result in important gains to currently used methods.

Keywords: Metagenomics, metataxonomics, 16S rRNA gene, taxonomic classification, graph matching

Resumo

As tecnologias de sequenciação paralela massiva têm contribuído imenso para os desenvolvimentos nas áreas que estudam amostras biológicas complexas, como a metataxonómica e a metagenómica. Um dos passos fundamentais na análise de dados derivados de amostras complexas é a determinação da origem taxonómica dos fragmentos genómicos sequenciados. Apesar da diversidade de ferramentas existente, os métodos de identificação taxonómica aplicados produzem resultados variáveis e com um grau de fiabilidade que é difícil de verificar, especialmente a níveis taxonómicos mais baixos, como espécie. Neste estudo, três estratégias de classificação baseadas em correspondência exacta de *k-mers*, implementadas nos programas Kraken, CLARK e Centrifuge, que são frequentemente utilizadas para classificação de dados de metagenómica, foram aplicadas a dados simulados de metataxonómica, criados a partir de sequências do gene 16S rRNA presentes em bases de dados de referência, de forma a avaliar e comparar a eficiência dessas estratégias de classificação. Esse desempenho foi também comparado com uma estratégia de classificação tipicamente aplicada em processos de metataxonómica e baseada em alinhamento de sequências. Cada estratégia foi avaliada através da classificação de sequências de 14 subregiões do gene 16S rRNA e tendo como referência três bases de dados e versões condensadas dessas bases de dados. No total, as estratégias baseadas em correspondência exacta de *k-mers* foram avaliadas com um total de 60 combinações de classificador, base de dados e subregião e a estratégia baseada em alinhamento de sequências foi avaliada com 45 combinações das mesmas variáveis. De entre as estratégias baseadas em correspondência exacta de *k-mers*, a estratégia baseada em correspodência de *k-mers* discriminativos destacou-se pelo desempenho superior que demonstrou ao classificar as sequências nos níveis taxonómicos de família, género e espécie. A classificação de sequências ao nível de família através de *k-mers* discriminativos foi, em média, 8.9% (exactidão de 0.878) mais exacta do que a estratégia baseada em alinhamentos de sequências (0.789) e ao nível taxonómico de género ambas as estratégias apresentaram um valor de exactidão igual (0.704). Ao nível de espécie, a classificação baseada em correspondência de *k-mers* discriminativos apresentou um desempenho superior às restantes estratégias, atingindo um valor médio de exactidão de aproximadamente 0.908 e superior a 0.846 para os 25% de valores de exactidão mais altos. Em termos comparativos, para a segunda melhor estratégia, também baseada em *k-mers*, os 25% de valores de

exactidão mais elevados variaram entre 0.332-0.705. De acordo com os valores absolutos de exactidão, as sub-regiões V4/V3-V4, V4/V3-V4 e V2/V1-V2 foram as que permitiram atingir os valores mais altos de exactidão aos níveis taxonómicos de familia, género e espécie, respectivamente.

De forma a explorar novos métodos que possam contribuir para melhorias na área de classificação taxonómica, a estratégia que apresentou melhor desempenho na avaliação das estratégias de classificação, baseada em *k-mers* discriminativos, foi utilizada para classificar sequências das sub-regiões V2-V4 e V6-V8. Os termos taxonómicos partilhados por sequências de sub-regiões diferentes foram utilizados para modelar as relações taxonómicas em grafos. Dois algoritmos de acoplamento (máximo peso e algoritmo húngaro) e um algoritmo que determina correspondências de muitos para muitos, com base na maioria de termos partilhados, foram utilizados para determinar correspondências entre os nós de cada grafo bipartido, que maximizassem o número de termos taxonómicos comuns entre os nós de cada correspondência determinada. A aplicação de dois valores de *cutoff* a um subconjunto das correspondências que capturavam a maior parte das diferenças taxonómicas previstas pelo classificador provocou uma diminuição do valor médio de exactidão para a sub-região V2-V4, ao nível de espécie, entre 0.73% e 1.99%, dependendo do algoritmo utilizado para determinar as correspondências. O mesmo procedimento levou a um aumento do valor médio da exactidão de 2.48%, ao nível de espécie, para a sub-região V6-V8, se as correspondências fossem determinadas sempre com base nos casos que partilhavam o maior número de termos taxonómicos. A aplicação dos valores de *cutoff* a correspondências determinadas pela regra de maioria de termos partilhados permitiu detectar sete espécies que não tinham sido detectadas através das previsões iniciais, diminuindo o número de espécies não detectadas de 31 para 24 (num total de 111 espécies). Uma abordagem alternativa consistiu na reclassificação das sequências representadas pelos nós das correspondências que apresentavam incompatibilidades através do BLAST. A classificação deste subconjunto de sequências permitiu melhorar o valor médio de exactidão, ao nível de espécie para a sub-região V2-V4, entre 1.12% e 1.53%, dependendo do algoritmo a partir do qual as correspondências tinham sido determinadas. Os valores de exactidão para a sub-região V6-V8 aumentaram entre 5.58% e 7.13%. O valor de *bitscore* calculado para cada previsão do BLAST foi utilizado para alterar os termos taxonómicos associados a cada nó, de forma a favorecer as classificações previstas com um *bitscore* mais alto. Esta estratégia levou a uma ligeira diminuição no valor de exactidão para ambas as sub-regiões. A reclassificação de sequências com o BLAST reduziu o número de espécies não detectadas, de 31 para um

mínimo de 13 ou 16, dependendo se as correspondências eram determinadas através do algoritmo húngaro ou da regra de maioria de termos partilhados, respectivamente. A alteração da informação taxonómica por favorecimento de valores de *bitscore* mais elevados permitiu detectar mais duas espécies, quando aplicado em correspondências determinadas através do algoritmo húngaro, e mais seis espécies, quando as correspondências tinham sido determinadas pela regra de maioria de termos partilhados. A combinação de sequências representadas por nós acoplados pelo algoritmo húngaro ou pela regra de maioria de termos partilhados levou a grupos *multilocus* em que as sequências constituintes partilhavam todos os termos taxonómicos com uma exactidão de 0.925 e 0.937, respectivamente.

Os resultados apresentados nesta dissertação demonstram que uma estratégia baseada em *k-mers* e desenvolvida para classificação de dados de metagenómica pode atingir níveis de desempenho considerados superiores em relação às estratégias testadas neste e noutros estudos para a classificação de dados de metataxonómica. Apesar de existirem algumas limitações inerentes à estratégia aqui aplicada, o nível de desempenho observado foi positivo e demonstra que existe potencial nesta abordagem para melhoramentos adicionais. A classificação de sequências de dois *loci* através de dois classificadores com estratégias diferentes, e a combinação dos resultados de ambos, evidencia as vantagens da combinação das previsões de mais do que um classificador, para minimizar as falhas da aplicação de uma única estratégia de classificação. Para além disso, a combinação dos resultados poderá representar melhorias em relação às estratégias actualmente aplicadas se forem desenvolvidas pontuações estatisticamente mais robustas que indiquem, com elevado nível de confiança, quais as previsões mais correctas, de forma a capturar os casos em que cada sub-região apresenta maior resolução taxonómica e promover a correcta alteração das classificações taxonómicas através de algoritmos de acoplamento. Em conclusão, este trabalho constitui a base para o possível desenvolvimento de uma estratégia que poderá trazer melhoramentos aos métodos de classificação actualmente utilizados.

Palavras Chave: Metagenómica, metataxonómica, gene 16S rRNA, classificação taxonómica, acoplamento